

UVTM: Universal Vehicle Trajectory Modeling with ST Feature Domain Generation

Yan Lin, Jilin Hu, Shengnan Guo, Bin Yang, Christian S. Jensen, Youfang Lin, Huaiyu Wan

Abstract—Vehicle movement is frequently captured in the form of trajectories, i.e., sequences of timestamped locations. Numerous methods exist that target different tasks involving trajectories such as travel-time estimation, trajectory recovery, and trajectory prediction. However, most methods target only one specific task and cannot be applied universally. Existing efforts to create a universal trajectory model often involve adding prediction modules for adapting to different tasks, while also struggle with incomplete or sparse trajectories.

To address these shortcomings, we propose the Universal Vehicle Trajectory Model (UVTM) designed to support different tasks based on incomplete or sparse trajectories without the need for retraining or extra prediction modules. To addresses task adaptability on incomplete trajectories, UVTM divide the spatio-temporal features of trajectories into three distinct domains. Each domain can be masked and generated independently to suit the input and output needs of specific tasks. To handle sparse trajectories effectively, UVTM is pre-trained by reconstructing densely sampled trajectories from sparsely sampled ones, allowing it to extract detailed spatio-temporal information from sparse trajectories. Experiments involving three representative trajectory-related tasks on two real-world vehicle trajectory datasets provide insight into the intended properties performance of UVTM and offer evidence that UVTM is capable of meeting its objectives.

Index Terms—Spatio-temporal data mining, vehicle trajectory mining, pre-training and fine-tuning, self-supervised learning.

arXiv:2402.07232v3 [cs.LG] 23 Apr 2024

1 INTRODUCTION

A trajectory is a sequence of timestamped locations that captures the movement of an object. Figure 1(a) provides an example where a trajectory \mathcal{T} captures the travel of a vehicle from l_1 to l_5 . This trajectory consists five timestamped GPS points: $\mathcal{T} = \langle (l_1, t_1), (l_2, t_2), \dots, (l_5, t_5) \rangle$. Substantial spatio-temporal information can be mined from this trajectory, e.g., 1) the travel time of the trip, $t_5 - t_1$; 2) the road segments that the trip visited, e_1, e_2, e_4 , and e_6 ; 3) the average travel speed on each road segment, which can reflect traffic conditions; 4) accelerations and decelerations that capture driving behavior. Such information provides a rich foundation for analyzing the movement patterns of individual vehicles and traffic patterns on road networks, which in turn powers various important tasks in Intelligent Transportation System (ITS). These tasks include trajectory prediction [1], [2], [3], [4], travel time estimation [5], [6], [7], [8], anomaly detection [9], [10], trajectory clustering [11], [12], [13], [14], and trajectory recovery [15], [16], [17].

Given the rich spatio-temporal information that trajectories can provide and their applicability across multiple tasks, the same set of trajectories is often utilized for various tasks simultaneously. Traditionally, individual models

are created and trained for each specific task. Thus, these models often solely concentrate on task-specific information, which hinders their adaptability to other tasks and potentially restricts their effectiveness. Moreover, training and storing multiple models for different tasks adversely affect computational and storage efficiency. Consequently, there is a pressing need to develop a universal trajectory model that can be trained once on a dataset and effectively address various types of tasks.

Several existing studies [18], [19], [20], [21] have introduced universal trajectory models through representation learning techniques. These methods commonly involve training a trajectory encoder, which maps each trajectory to an embedding vector. However, utilizing the embedding vector alone is insufficient for most tasks. To make task-specific predictions, supplementary prediction modules must be connected to the embedding vector — as illustrated in Figure 1(a) — which requires additional fine-tuning processes. Apart from the need for fine-tuning, existing methods suffer from two significant shortcomings that prevent them from fully achieving the objective of constructing a general trajectory model.

First, existing methods face limitations in adapting to different types of tasks due to the requirement of the integrity of a trajectory's features. There is a wide range of trajectory-related tasks that involve varying arrangements of input and output features. Some tasks, especially, involve incomplete trajectory points and spatio-temporal features as the input. For instance, in origin-destination (OD) travel time estimation [22], only the origin, destination, and departure time of a trajectory are known prior to its occurrence. Similarly, in trajectory recovery [23], there are often long spans of missing trajectory points among the known ones.

Existing methods typically have strict requirements re-

Corresponding author: Huaiyu Wan

• Yan Lin, Shengnan Guo, Youfang Lin, and Huaiyu Wan are with the Beijing Key Laboratory of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China, and the Key Laboratory of Intelligent Passenger Service of Civil Aviation, CAAC, Beijing, 101318, China. Jilin Hu and Bin Yang are with the School of Data Science and Engineering, East China Normal University, Shanghai 200050, China. Christian S. Jensen is with the Department of Computer Science, Aalborg University, Aalborg, 9220, Denmark.
E-mail: ylincs@bjtu.edu.cn; jlhu@dase.ecnu.edu.cn; guoshm@bjtu.edu.cn; byang@dase.ecnu.edu.cn; csj@cs.aau.dk; yflin@bjtu.edu.cn; hywan@bjtu.edu.cn.

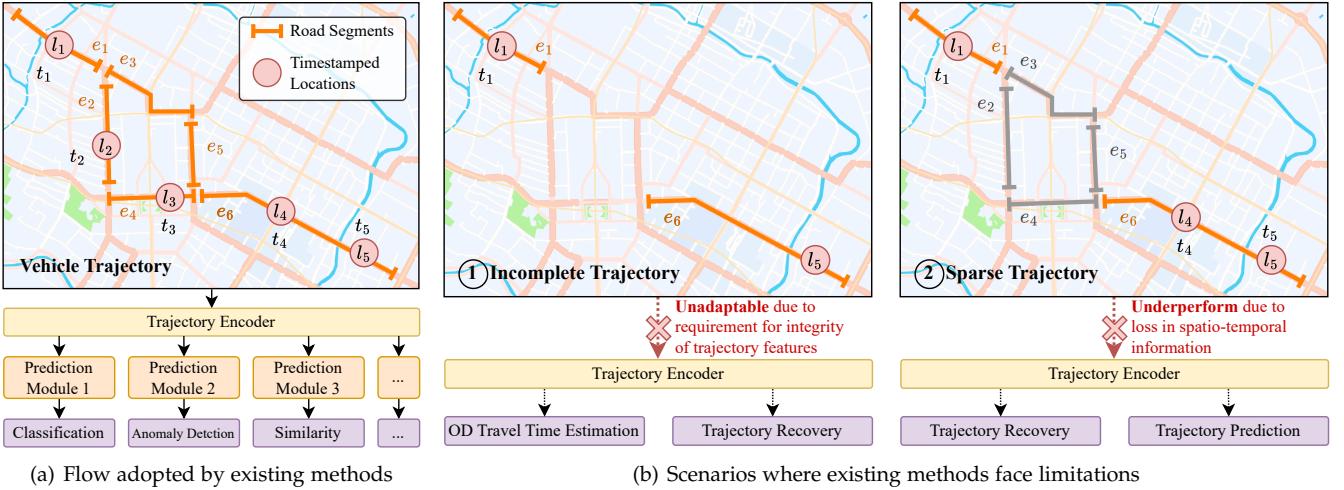


Fig. 1. Overview of the existing efforts on universal vehicle trajectory modeling and their limitations.

garding the integrity of the input features in order to function properly. Specifically, they commonly specify that the spatio-temporal features of all points in a trajectory be provided as input. For example, t2vec [18] requires the coordinate and timestamp of each trajectory point, while Trembr [19] and START [20] necessitate the road segment and timestamp for each point. Consequently, these methods are unable to adapt to tasks such as OD travel time estimation and trajectory recovery, where the requirement of the integrity of a trajectory's features cannot be fulfilled, as illustrated in Figure 1(b), scenario ①.

Second, existing methods scale poorly with long vehicle trajectories and underperform on re-sampled sparse trajectories. In real-world settings, vehicle locations are often sampled at high rates, resulting in long vehicle trajectories. State-of-the-art methods, including PreCLN [24] and START [20], utilize Transformer-based [25] trajectory encoders to capture long-term correlations in these trajectories. However, Transformers exhibit poor scalability with long trajectories due to their quadratic complexity, $O(N^2)$, in the trajectory length N . Additionally, the data storage requirements for these lengthy trajectories are substantial.

To address these issues, a common approach is to re-sample the original trajectories into sparser ones with larger sampling intervals to enhance computational and storage efficiency. However, existing methods encounter performance issues when working with such sparse trajectories as they struggle to address the information loss caused by the sparsity properly. For instance, in Figure 1(b), scenario ②, a sparse trajectory $\mathcal{T}' = \{(l_1, t_1), (l_4, t_4), (l_5, t_5)\}$ is re-sampled from the original trajectory \mathcal{T} . When existing methods are fed with \mathcal{T}' as input, they tend to overlook spatio-temporal information not explicitly captured in \mathcal{T}' . For example, distinguishing whether the vehicle traverses road segments e_2, e_4 or e_3, e_5 between points (l_1, t_1) and (l_4, t_4) poses a challenge for existing methods.

To this end, we propose the *Universal Vehicle Trajectory Model* (UVTM). UVTM is designed to be trained once and then adjusted to effectively address various trajectory-related tasks, facilitating its adaptability and efficiency in ITS applications. UVTM incorporates specialized designs

that excel at overcoming the limitations of current methods mentioned earlier. To address the first limitation, we divide the spatio-temporal features in trajectories into three distinct domains. Each domain can be masked and generated independently to meet the specific input and output requirements of the task at hand. These designs empower UVTM to adapt to tasks involving incomplete trajectories. In response to the second limitation, we pre-train UVTM to enable it to extract spatio-temporal and road segment features embedded in the densely sampled trajectories, even when presented with only the sparsely re-sampled ones. This pre-training ensures that UVTM consistently delivers strong performance when confronted with sparse trajectory data. In summary, the primary contributions of the paper are as follows.

- We propose the UVTM, a versatile and robust universal vehicle trajectory model. UVTM can be trained once and effectively adapt to multiple trajectory-related tasks without additional prediction modules, even when confronted with incomplete or sparse trajectories.
- We separate the features in trajectories into three domains, allowing each domain to be masked and generated separately based on the specific input and output requirements of the task. This flexibility enables the UVTM to adapt to various types of tasks, particularly those involving incomplete trajectories.
- We pre-train UVTM by reconstructing densely sampled trajectories using their re-sampled sparse counterparts. This process enhances UVTM's robustness to sparsity and maintains its performance on sparse trajectories.
- We conduct extensive experimental study on two real-world datasets and three representative tasks, offering insight into the performance properties of the proposed model, and providing evidence that the model is capable of meeting its design goals.

2 RELATED WORK

2.1 Task-specific Trajectory Models

Task-specific models are designed and trained for specific tasks. These models are commonly used for trajectory-

related tasks due to their ease of implementation.

For travel time estimation, path-based approaches including WDR [26], DeepTTE [27], DeepETA [28], WD-DRA [29], and DRTTE [30] predict travel times based on the given travel paths. On the other hand, origin-destination-based approaches including TEMP [31], ST-NN [32], MURAT [33], DeepOD [5], and DOT [22] estimate travel times by considering the origin, destination, and departure time. For trajectory recovery, TrImpute [23] is a non-learning-based method. In contrast, AttnMove [15], MTrajRec [16], and RNTrajRec [17] are learning-based methods built on the seq2seq framework [34]. For trajectory prediction, models like DeepMove [1], HST-LSTM [2], and ACN [35] leverage recurrent neural networks [36] to capture sequential patterns in trajectories. PreCLN [37] integrates contrastive learning [38] to improve prediction accuracy.

While task-specific trajectory models have their advantages, their adaptability and computational efficiency are limited in ITS applications. Models tailored to specific tasks cannot be easily repurposed for other tasks, leading to the need for separate models for each task, which can adversely affect computational resources and storage efficiency.

2.2 Universal Trajectory Models

In response to the limitations of task-specific models, there is a growing interest in universal trajectory models that can accommodate multiple tasks.

Among these models, trajectory2vec [39] constructs behavior sequences from trajectories to extract key information and then utilizes an auto-encoding framework [40] to compress each sequence into an embedding vector. t2vec [18] employs a denoising auto-encoding framework to enhance its resilience to trajectory noise. Trembr [19] leverages auto-encoding techniques to extract road network and temporal information embedded in trajectories effectively. SML [24] integrates the contrastive predictive coding framework [41], providing an innovative method for learning embedding vectors for trajectories. START [20] introduces a comprehensive approach to trajectory embedding learning by combining masked language model [42] and SimCLR [38] to enhance its learning capability.

While these universal trajectory models exhibit versatility, they often require additional prediction modules to generate task-specific predictions based on their output embedding vectors. Furthermore, their adaptability is limited, particularly for tasks that include incomplete trajectories, and they tend to underperform when dealing with sparse trajectories.

3 PRELIMINARIES

3.1 Definitions

Definition 1 (Road Network). A road network is modeled as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where \mathcal{V} is a set of nodes, each node $v_i \in \mathcal{V}$ models an intersection between road segments or the end of a segment, and \mathcal{E} is a set of edges, each edge $e_i \in \mathcal{E}$ models a road segment linking two nodes. An edge is given by starting and ending nodes: $e_i = (v_j, v_k)$.

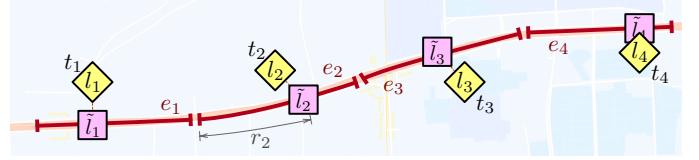


Fig. 2. A trajectory and its map-matched counterpart.

Definition 2 (Trajectory). A trajectory \mathcal{T} is a sequence of timestamped point locations: $\mathcal{T} = \langle (l_1, t_1), (l_2, t_2), \dots, (l_{|\mathcal{T}|}, t_{|\mathcal{T}|}) \rangle$, where $l_i = (l_i^{\text{long}}, l_i^{\text{lat}})$ are the spatial coordinates of the i -th location, l_i^{long} and l_i^{lat} denote longitude and latitude, respectively, and timestamp t_i is the time at which l_i is visited.

Definition 3 (Sampling Interval). A trajectory's sampling interval η is the time interval between its consecutive points, i.e., $\eta = t_i - t_{i-1}, i \in \{2, 3, \dots, |\mathcal{T}|\}$. Given a trajectory that is originally densely sampled, its sparse counterpart can be obtained by re-sampling the trajectory with a larger interval μ .

It is important to note that we assume consistent sampling intervals to ensure consistency in the experimental evaluation. Nevertheless, the proposed model can be applied to trajectories with varying sampling intervals.

Definition 4 (Map-matched Trajectory). By using a map-matching algorithm [43], a trajectory \mathcal{T} can be projected onto the underlying road network \mathcal{G} . The map-matched trajectory can be denoted as $\tilde{\mathcal{T}} = \langle (\tilde{l}_1, t_1), (\tilde{l}_2, t_2), \dots, (\tilde{l}_{|\mathcal{T}|}, t_{|\mathcal{T}|}) \rangle$, where $\tilde{l}_i = (e_i, r_i)$, $e_i \in \mathcal{E}$ is the map-matched road segment occupied by location l_i , and r_i is the fraction of the length of the road segment traveled by time t_i .

Example 1. Figure 2 gives an example of a trajectory $\mathcal{T} = \langle (l_1, t_1), (l_2, t_2), (l_3, t_3), (l_4, t_4) \rangle$, where the locations are denoted in diamonds. Map-matching yields map-matched trajectory $\tilde{\mathcal{T}} = \langle (\tilde{l}_1, t_1), (\tilde{l}_2, t_2), (\tilde{l}_3, t_3), (\tilde{l}_4, t_4) \rangle$, where the locations are in squares. For example, l_2 is map-matched onto \tilde{l}_2 , which is in the middle of road segment e_2 . Thus, we use r_2 to denote the fraction of the length of e_2 that the vehicle has traveled, resulting in $\tilde{l}_2 = (e_2, r_2)$.

3.2 Problem Statement

Universal Vehicle Trajectory Modeling. The objective is to construct a universal vehicle trajectory model f_θ , where θ denotes a set of learnable parameters. During evaluation, f_θ takes a certain arrangement of trajectory \mathcal{T} as input and generates an output tailored to the particular task at hand, denoted as $\hat{Y} = f_\theta(\text{arrange}(\mathcal{T}))$. For example, for OD travel time estimation, $\text{arrange}(\mathcal{T})$ extracts the origin, destination, and departure time of \mathcal{T} , and \hat{Y} is the estimated travel time; for trajectory prediction, $\text{arrange}(\mathcal{T})$ retains the historical part of \mathcal{T} , and \hat{Y} is the predicted future part of \mathcal{T} .

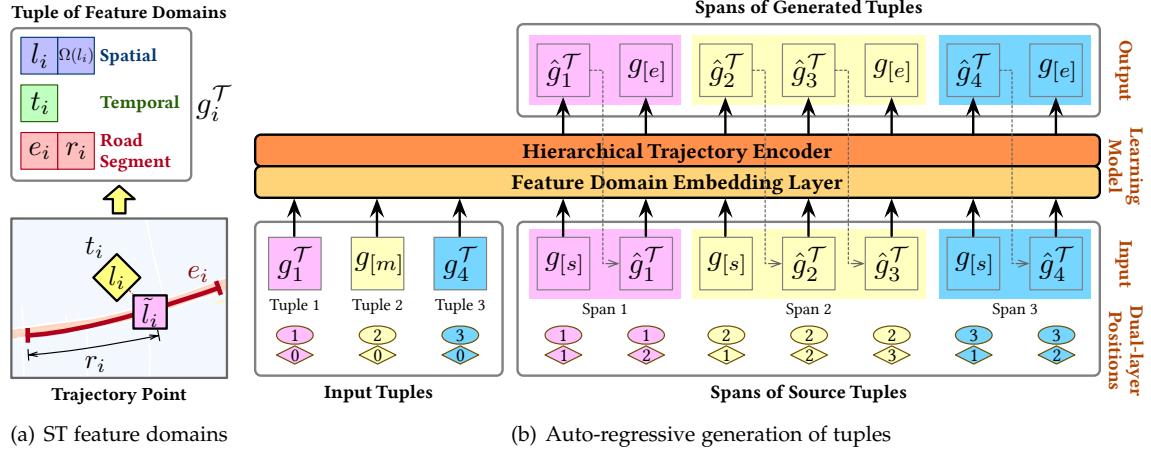


Fig. 3. Overall framework of the proposed Universal Vehicle Trajectory Model (UVTM).

4 METHODOLOGY

4.1 Overall Framework

The overall framework of the proposed UVTM is illustrated in Figure 3. We first divide the spatio-temporal features of trajectories into three distinct domains: spatial, temporal, and road segment, as depicted in Figure 3(a). Each point in the trajectory can be represented as a *tuple* consisting of these three domains. We refer to the tuple representing the i -th point in the trajectory \mathcal{T} as $g_i^{\mathcal{T}}$.

Each feature domain in $g_i^{\mathcal{T}}$ can be masked and generated independently. Specifically, a domain can be replaced with a special mask token to indicate that it needs to be generated. Additionally, the tuple $g[m]$, filled with the mask token, can represent all feature domains of a sub-trajectory that require generation. For each input tuple containing the mask token, a series of tuples is auto-regressively generated, creating a *span* of generated tuples. This generation process begins with a tuple $g[s]$ filled with a special start token added to the input, and proceeds iteratively by adding the last generated tuple to the input to produce the next generated tuple. The process concludes when a tuple $g[e]$ filled with a special end token is produced. Figure 3(b) provides an example where three spans are generated corresponding to the three input tuples. The flexibility provided by the above design enables UVTM to adapt to various tasks, as it can accept task-specific input arrangements and generate the desired output.

To facilitate the model's ability to extract information and correlations from feature domains, we introduce the feature domain embedding layer and the hierarchical trajectory encoder. Additionally, to enhance the model's robustness to sparse trajectories, the learnable parameters in the model are pre-trained by reconstructing the feature domains of densely sampled trajectories given their re-sampled sparse counterparts. The next sections detail designs and modules in the proposed UVTM.

4.2 ST Feature Domains

In order to improve the model's ability to handle different tasks, we divide the spatio-temporal (ST) features in trajectories into three domains. Each domain can be masked and

generated separately, depending on the specific input and output requirement of the respective task.

4.2.1 Tuple of Feature Domains

Given the i -th point (l_i, t_i) of trajectory \mathcal{T} , we split its spatio-temporal features into three domains: spatial, temporal, and road segment.

The spatial domain contains features related to the coordinate l_i . These features include l_i itself, and a set $\Omega(l_i)$ of road segments located within a distance of δ meters from l_i . $\Omega(l_i)$ enhances the model's understanding of the relationship between coordinates and road segments.

The temporal domain contains the timestamp t_i measured in seconds. To standardize t_i , we subtract it from the initial timestamp t_0 of the trajectory.

The road segment domain consists of road segment e_i and fraction r_i from \tilde{l}_i , which can be obtained through map-matching the coordinate l_i following Definition 4.

Finally, we transform (l_i, t_i) into a *tuple* consisting of the above three domains as:

$$g_i^{\mathcal{T}} = ((l_i, \Omega(l_i)), t_i, (e_i, r_i)) \quad (1)$$

4.2.2 Auto-regressive Generation

To generate a feature domain in the tuple $g_i^{\mathcal{T}}$, we fill that domain with the mask token $[m]$, denoting a masked domain. This tuple is then provided to the model as input. Next, the masked domain is generated through an auto-regressive process, where each input tuple corresponds to a series of generated tuples referred to as a *span*. The generation begins by adding a tuple $g[s] = ([s], [s], [s])$ with all domains filled with the start token $[s]$ to the input. The model then outputs the generated tuple $\hat{g}_i^{\mathcal{T}}$. Subsequently, $\hat{g}_i^{\mathcal{T}}$ is added back to the input. Finally, the model outputs a tuple $g[e] = ([e], [e], [e])$ with all domains filled with the end token $[e]$, indicating the completion of the generation. An example of this generation process can be observed in the input tuple $g_1^{\mathcal{T}}$ and the generated span $\langle \hat{g}_1^{\mathcal{T}}, g[e] \rangle$ in Figure 3(b).

In cases where all feature domains of a sub-trajectory need to be generated, this sub-trajectory is replaced by a single tuple $g[m] = ([m], [m], [m])$ with all domains filled with the mask token $[m]$. The generation follows a similar

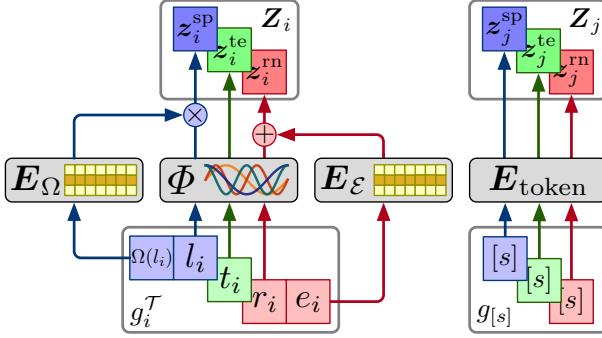


Fig. 4. The feature domain embedding layer.

auto-regressive process as described above, with the distinction that the generation process is iterated by appending the previously generated tuple to the input to derive the subsequent generated tuple. The end of the generation is determined by the model itself, i.e., the output of the tuple $g_{[e]}$. An example of this generation process can be observed in the input tuple $g_{[m]}$ and the generated span $\langle \hat{g}_2^T, \hat{g}_3^T, g_{[e]} \rangle$ in Figure 3(b). This design further enhances the model's flexibility, particularly in tasks such as trajectory recovery and trajectory prediction, where the length of the trajectory that needs to be generated cannot be predetermined.

When multiple input tuples with masked domains are provided, these tuples are organized into an input sequence for the model. Next, the model generates the spans sequentially, with each span being generated following the above described procedures. Figure 3(b) provides an example, where the three spans corresponding to the three input tuples are sequentially generated.

4.2.3 Dual-layer Positions

Since each input tuple g_i^T corresponds to a span of generated tuples, we follow the practice of GLM [44] and employ dual-layer positions.

The first layer of positions, denoted as \mathcal{P}_i^1 , indicates the relationship between each input tuple g_i^T and the span of generated tuples. Specifically, both g_i^T and its corresponding span of generated tuples have the same first layer positions, which is i .

The second layer of positions, denoted as \mathcal{P}_i^2 , represents the order of tuples within each span of generated tuples. For g_i^T , the second layer position is always 0. Conversely, for the span of generated tuples of length N , the positions are sequentially numbered starting from 1 up to N .

4.2.4 Feature Domain Embedding Layer

To enhance the model's ability to extract information from the three feature domains of each tuple, we propose an embedding layer that projects domains into the latent space, as illustrated in Figure 4.

The continuous features in the tuple g_i^T , i.e., l_i^{lng} , l_i^{lat} , t_i , and r_i , exhibit periodic characteristics. To mirror these characteristics, we encode these continuous features by drawing inspiration from the learnable Fourier features [45], [46], leveraging the cyclic nature of trigonometric functions. Given an input continuous feature $x \in \mathbb{R}$, the encoding

module Φ projects it onto a d -dimensional latent space, defined as follows:

$$\Phi(x) = \mathbf{W}_\Phi[\cos(x\mathbf{v}_\Phi) \parallel \sin(x\mathbf{v}_\Phi)], x \in \{l_i^{\text{lng}}, l_i^{\text{lat}}, t_i, r_i\}, \quad (2)$$

where $\mathbf{v}_\Phi \in \mathbb{R}^{d/2}$ represents a learnable projection vector, and $\mathbf{W}_\Phi \in \mathbb{R}^{d \times d}$ denotes a learnable projection matrix. Importantly, distinct sets of \mathbf{v}_Φ and \mathbf{W}_Φ are utilized for the four types of continuous features.

On the other hand, segment e_i , the segments in $\Omega(l_i)$, and the special tokens $\{[m], [s], [e]\}$ are discrete features. For these, we employ an index-fetching embedding module for each feature. These modules involve learnable matrices $\mathbf{E}_\mathcal{E} \in \mathbb{R}^{|\mathcal{E}| \times d}$, $\mathbf{E}_\Omega \in \mathbb{R}^{|\mathcal{E}| \times d}$, and $\mathbf{E}_{\text{token}} \in \mathbb{R}^{3 \times d}$. The embedding vector for a discrete feature is retrieved as the corresponding row vector from these matrices. For example, the embedding vector of road segment e_i is obtained as the e_i -th row vector from $\mathbf{E}_\mathcal{E}$, denoted as $\mathbf{E}_\mathcal{E}(e_i)$.

Finally, the above embedding vectors are gathered to form a embedding matrix for the tuple g_i^T . Formally, the embedding vectors for the spatial, temporal, and road network domains of g_i^T are calculated as follows:

$$\begin{aligned} \mathbf{z}'_i^{\text{sp}} &= \Phi(l_i^{\text{lng}}) + \Phi(l_i^{\text{lat}}) \\ \mathbf{z}_i^{\text{sp}} &= \mathbf{z}'_i^{\text{sp}} + \text{MultiHead}(\mathbf{z}'_i^{\text{sp}}, \mathbf{E}_\Omega(l_i), \mathbf{E}_\Omega(l_i)) \\ \mathbf{z}_i^{\text{te}} &= \Phi(t_i) \\ \mathbf{z}_i^{\text{rn}} &= \mathbf{E}_\mathcal{E}(e_i) + \Phi(r_i) \end{aligned} \quad (3)$$

Here, MultiHead represents the dot-product attention defined in the Transformer [25], with N_h attention heads. $\mathbf{E}_\Omega(l_i)$ denotes the set of embedding vectors for all segments in $\Omega(l_i)$. In scenarios where any feature domain is one of the special tokens, the embedding vector of that domain is the embedding vector of the corresponding special token. Finally, the embedding matrix for the tuple g_i^T is expressed as $\mathbf{Z}_i = \langle \mathbf{z}_i^{\text{sp}}, \mathbf{z}_i^{\text{te}}, \mathbf{z}_i^{\text{rn}} \rangle \in \mathbb{R}^{3 \times d}$.

4.3 Hierarchical Trajectory Encoder

To model the correlations between different tuples in a trajectory and to execute the generation process described in Section 4.2.2, we propose a trajectory encoder that takes a sequence of matrices \mathbf{Z}_i and generates a sequence of tuples \hat{g}_i^T as output.

4.3.1 Hierarchical Attention

To extract the correlations between the feature domains of a tuple, we take the embedding matrix \mathbf{Z}_i and apply a matrix-level self-attention mechanism. This mechanism, along with mean pooling and position encoding, allows us to obtain the hidden state \mathbf{h}_i of the tuple. Formally:

$$\mathbf{h}_i = \text{Mean}(\text{MultiHead}(\mathbf{Z}_i, \mathbf{Z}_i, \mathbf{Z}_i)) + \text{PE}(\mathcal{P}_i^1) + \text{PE}(\mathcal{P}_i^2), \quad (4)$$

where \mathbf{h}_i is the hidden state of the i -th tuple, \mathbf{Z}_i is treated as a sequence of length 3, and PE represents the positional encoding from the Transformer [25].

Performing the generation process described in Section 4.2.2 involves a sequence of tuples provided as input. We obtain the hidden state of each tuple in the sequence using Equation 4, resulting in a sequence $\mathbf{H} = \langle \mathbf{h}_1, \mathbf{h}_2, \dots \rangle$ of hidden states. Then, we extract the correlations between

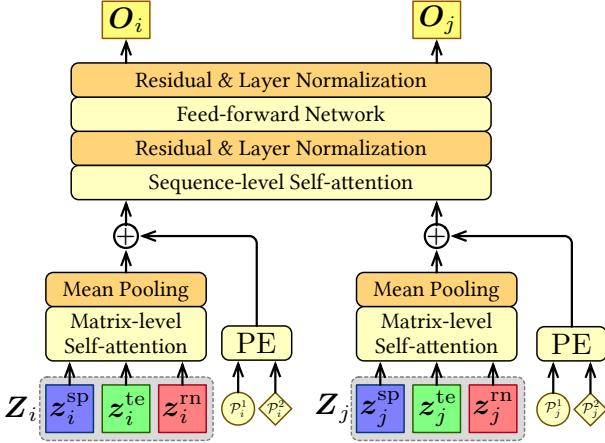


Fig. 5. The hierarchical trajectory encoder.

different tuples in the sequence using a sequence-level self-attention mechanism. This mechanism incorporates a residual connection, layer normalization, and a feed-forward network. Formally:

$$\begin{aligned} \mathbf{H}' &= \text{MultiHead}(\mathbf{H}, \mathbf{H}, \mathbf{H}) \\ \mathbf{O}' &= \text{LayerNorm}(\mathbf{H}' + \mathbf{H}) \\ \mathbf{O} &= \text{LayerNorm}(\text{FFN}(\mathbf{O}') + \mathbf{O}'), \end{aligned} \quad (5)$$

where \mathbf{O} represents the sequence of output states and FFN is the feed-forward network composed of two fully connected layers.

4.3.2 Output Layer

To generate the tuple of feature domains from the i -th output state \mathbf{O}_i , we utilize four fully connected networks to establish an output layer. Formally:

$$\begin{aligned} \hat{l}_i &= \mathbf{W}_c \mathbf{O}_i + \mathbf{b}_c \\ \hat{t}_i &= \mathbf{W}_t \mathbf{O}_i + \mathbf{b}_t \\ \hat{e}_i &= \text{argmax}(p(\hat{e}_i)), p(\hat{e}_i) = \text{Softmax}(\mathbf{W}_e \mathbf{O}_i + \mathbf{b}_e) \\ \hat{r}_i &= \mathbf{W}_r \mathbf{O}_i + \mathbf{b}_r, \end{aligned} \quad (6)$$

where $\mathbf{W}_c \in \mathbb{R}^{2 \times d}$, $\mathbf{W}_t \in \mathbb{R}^{1 \times d}$, $\mathbf{W}_e \in \mathbb{R}^{(|\mathcal{E}|+1) \times d}$, and $\mathbf{W}_r \in \mathbb{R}^{1 \times d}$ denote the projection matrices, while $\mathbf{b}_c \in \mathbb{R}^2$, $\mathbf{b}_t \in \mathbb{R}$, $\mathbf{b}_e \in \mathbb{R}^{(|\mathcal{E}|+1)}$, and $\mathbf{b}_r \in \mathbb{R}$ are biases. The generated tuple is then formed as:

$$\hat{g}_i^{\mathcal{T}} = ((\hat{l}_i, \Omega(\hat{l}_i)), \hat{t}_i, (\hat{e}_i, \hat{r}_i)) \quad (7)$$

If $\hat{e}_i = [e]$, we denote $\hat{g}_i^{\mathcal{T}} = g_{[e]}$.

Given the ground truth tuple $g_i^{\mathcal{T}}$, we can supervise the generated tuple $\hat{g}_i^{\mathcal{T}}$ with the following loss function:

$$\mathcal{L}_i = \mathbb{1}_i^{[e]} \left(\frac{1}{2} \|\hat{l}_i - l_i\|_2 + \|\hat{t}_i - t_i\|_2 + \|\hat{r}_i - r_i\|_2 \right) - \log p(\hat{e}_i), \quad (8)$$

where $\mathbb{1}_i^{[e]}$ equals 0 if $g_i^{\mathcal{T}} = g_{[e]}$ and equals 1 otherwise.

4.4 Pre-training

In order to improve the model's ability to handle sparsity, we propose pre-training the model by reconstructing feature domains in densely sampled trajectories based on their re-sampled sparse counterparts. This approach allows the model to extract detailed spatio-temporal and road segment

information from dense trajectories, even when only sparse trajectories are given. Consequently, the model can maintain its performance when dealing with sparse trajectories that have varying sampling intervals.

4.4.1 Reconstruction Procedure

Initially, a dense trajectory \mathcal{T} is selected from the dataset, where its sampling interval η is no longer than 15 seconds. The Fast Map Matching (FMM) algorithm [47] is then applied to obtain the trajectory's map-matched counterpart $\tilde{\mathcal{T}}$. Simultaneously, to emulate sparse trajectories with longer sampling intervals, \mathcal{T} is resampled at an interval μ , where $\mu > \eta$ and μ is divisible by η , to derive a sparse counterpart \mathcal{T}' . This process is formulated as follows:

$$\mathcal{T}' = \langle (l_1, t_1), (l_{1+\mu/\eta}, t_{1+\mu/\eta}), (l_{1+2\mu/\eta}, t_{1+2\mu/\eta}), \dots, (l_{|\mathcal{T}'|}, t_{|\mathcal{T}'|}) \rangle \quad (9)$$

Notice that the last trajectory point in \mathcal{T} is retained in \mathcal{T}' to maintain the integrity of the re-sampled trajectory.

Comparing the trajectory point (l_i, t_i) in \mathcal{T}' to the corresponding point (\tilde{l}_i, t_i) in $\tilde{\mathcal{T}}$, the road segment domain is absent and needs to be generated. This can be represented by a tuple $g_i^{\mathcal{T}'} = ((l_i, \Omega(l_i)), t_i, [m])$. On the other hand, the sub-trajectory of \mathcal{T} that lies between each pair of consecutive points in \mathcal{T}' also requires generation. This can be denoted by the tuple $g_{[m]}$.

The objective is to reconstruct the missing feature domains in \mathcal{T}' , which involves generating all feature domains in \mathcal{T} . To achieve this, a sequence:

$$\text{input} = \langle g_1^{\mathcal{T}'}, g_{[m]}, g_{1+\mu/\eta}, g_{[m]}, g_{1+2\mu/\eta}, \dots, g_{|\mathcal{T}'|}^{\mathcal{T}'} \rangle \quad (10)$$

consisting of tuples mentioned above is provided as the input sequence to the model. Next, the model generates a span for each tuple in the input sequence, following the process described in Section 4.2.2. Specifically, for the input tuple $g_i^{\mathcal{T}'}$, the model generates the span $\langle \hat{g}_i^{\mathcal{T}'}, g_{[e]} \rangle$. Conversely, for the input tuple $g_{[m]}$, the model generates the span representing the sub-trajectory in \mathcal{T} occurring between the pair of consecutive points in \mathcal{T}' , with the tuple $g_{[e]}$ at the end.

To allow the model to extract bi-directional correlations from \mathcal{T} , we generate the spans of tuples with a shuffled order of the spans [48]. The shuffled spans are then generated sequentially, with the head and tail of each span concatenated together as a sequence. For example, the sequence:

$$\text{output} = \langle \hat{g}_j^{\mathcal{T}'}, \hat{g}_{j+1}^{\mathcal{T}'}, g_{[e]}, \hat{g}_i^{\mathcal{T}'}, \hat{g}_{i+1}^{\mathcal{T}'}, g_{[e]} \rangle \quad (11)$$

consisting of spans of generated tuples is the output of the model for one pre-training sample, with the generation order of $\langle \hat{g}_j^{\mathcal{T}'}, \hat{g}_{j+1}^{\mathcal{T}'}, g_{[e]} \rangle$ first, followed by $\langle \hat{g}_i^{\mathcal{T}'}, \hat{g}_{i+1}^{\mathcal{T}'}, g_{[e]} \rangle$. Additionally, we apply a causal mask on the MultiHead in Equation 5 to prevent information leakage.

4.4.2 Optimization Objective

Given a dense trajectory dataset \mathbb{T} , where the sampling interval of each trajectory is below 15 seconds, we employ a variety of re-sampling intervals to create multiple sparse trajectories from each dense trajectory. This process enhances the model's ability to handle trajectories with varying sampling intervals. During pre-training, we use three re-sample intervals: 1 minute, 2 minutes, and 4 minutes.

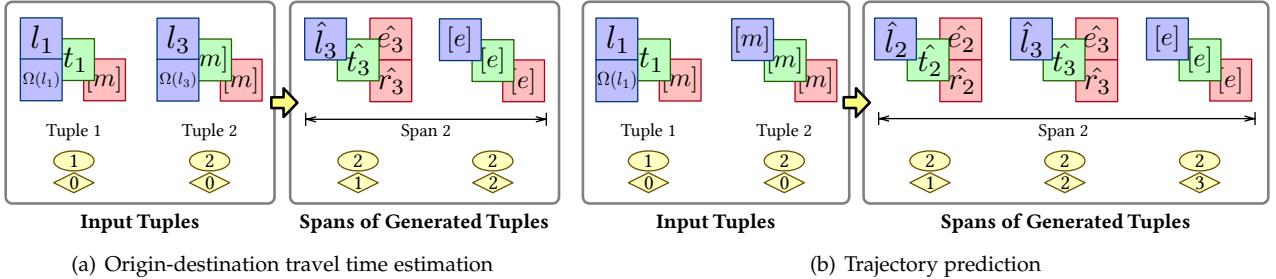


Fig. 6. Unique input and output arrangements for different tasks.

Each pair of dense trajectory \mathcal{T} and its sparsely resampled counterpart \mathcal{T}' serves as a pre-training sample in Section 4.4.1. By utilizing the loss function in Equation 8, the loss function $\mathcal{L}_{\mathcal{T}}$ of each pre-training sample can be calculated as the average of the loss \mathcal{L}_i for all generated tuples corresponding to the dense trajectory \mathcal{T} . Finally, during the pre-training process, the optimization objective is to minimize the sum of the loss $\mathcal{L}_{\mathcal{T}}$ across all pre-training samples constructed earlier. Formally:

$$\operatorname{argmax}_{\theta} \sum_{\mu \in \{1\text{min}, 2\text{min}, 4\text{min}\}} \sum_{\mathcal{T} \in \mathbb{T}} -\mathcal{L}_{\mathcal{T}}, \quad (12)$$

where θ denotes the set of learnable parameters in the model.

4.5 Task Adaptation

The model's flexibility allows it to adapt to various trajectory-related tasks after pre-training, especially those involving input of incomplete trajectories. We provide three representative tasks as examples.

4.5.1 Origin-Destination Travel Time Estimation (OD TTE)

We adopt the OD-based variant of the TTE task, which involves estimating the travel time of a trajectory given only its origin l_o , destination l_d , and the departure time t_o . The model is provided with an input sequence $\langle (l_o, t_o, [m]), (l_d, [m], [m]) \rangle$. Next, the model generates the span corresponding to the input tuple $(l_d, [m], [m])$ following the procedure outlined in Section 4.2.2. The generated span is expressed as $\langle (\hat{l}_d, \Omega(\hat{l}_d)), \hat{t}_d, (\hat{e}_d, \hat{r}_d), g_{[e]} \rangle$. Finally, \hat{t}_d is regarded as the estimated arrival time. Figure 6(a) provides an example of the above procedure.

4.5.2 Trajectory Recovery (TR)

This task aims to recover a sparse trajectory's dense counterpart. The proposed model can handle this task using a similar procedure as described in Section 4.4.1. The only difference is that the spans can be generated in their original order without shuffling. Whether a sub-trajectory between a pair of sparse trajectory's consecutive points is absent can be deduced from the sampling interval. Specifically, suppose we have two trajectory points (l_i, t_i) and (l_{i+1}, t_{i+1}) from a sparse trajectory, and a desired sampling interval η for the recovered dense trajectory. If $t_{i+1} - t_i > \eta$ then we insert the tuple $g_{[m]}$ between these two points, indicating that there is an absent sub-trajectory to be generated.

4.5.3 Trajectory Prediction (TP)

This task aims to predict future trajectory points given historical ones. Suppose the first n points of a trajectory \mathcal{T} are known. The model is provided with an input sequence $\langle g_1^{\mathcal{T}}, \dots, g_n^{\mathcal{T}}, g_{[m]} \rangle$. Next, the model generates the span of future tuples corresponding to $g_{[m]}$ following the procedure outlined in Section 4.2.2. The generated span can be expressed as $\langle \hat{g}_{n+1}^{\mathcal{T}}, \hat{g}_{n+2}^{\mathcal{T}}, \dots, \hat{g}_N^{\mathcal{T}}, g_{[e]} \rangle$, where N is the length of the trajectory. An example of the above procedure is illustrated in Figure 6(b).

The model can be directly applied to a task post pre-training in a zero-shot manner by providing the model with the task-specific input arrangements detailed above. In our experiments, fine-tuning with task-specific generations is conducted post pre-training to further improve the model's performance on tasks. Subsequent sections will delve into the comparison of performance between fine-tuned and non-fine-tuned models, the rate of convergence in fine-tuning, and the scalability of the fine-tuning dataset.

5 EXPERIMENTS

To access the performance of the proposed model, we conduct experiments on two real-world trajectory datasets under varying experimental settings.

5.1 Datasets

We conduct experiments using two real-world vehicle trajectory datasets: Chengdu and Porto. These datasets consist of vehicle trajectories from Chengdu, China, and Porto, Portugal, respectively. The Chengdu dataset was released by Didi¹ and consists of GPS trajectories of taxis operating in Chengdu. The Porto dataset was released on Kaggle² for a vehicle trajectory prediction contest. We also fetch the road network of Chengdu and Porto from OpenStreetMap³ for map-matching. For consistency in our analyses, these datasets are both standardized to a sampling interval of 15 seconds. Trajectories that feature fewer than 6 points, being relatively short, are excluded from our study. An overview of dataset statistics can be found in Table 1.

1. <https://gaia.didichuxing.com/>
2. <https://www.kaggle.com/competitions/pkdd-15-predict-taxi-service-trajectory-i/data>
3. <https://www.openstreetmap.org/>

TABLE 1
Dataset statistics.

Dataset	Chengdu	Porto
Time span	10.01–11.30, 2018	07.01–09.01, 2013
Longitude scope	104.04300~104.12654	-8.65200~−8.57801
Latitude scope	30.65523~30.72699	41.14201~41.17399
Number of road segments	2,505	2,225
Number of trajectories	121,394	55,120
Number of points	3,032,212	1,482,751

5.2 Comparison Methods

To assess the performance of the proposed model across the three trajectory-related tasks outlined in Section 4.5, we benchmark its performance against a variety of state-of-the-art methods. These include both universal trajectory models and methods specifically designed for the tasks at hand.

5.2.1 OD TTE Methods

We benchmark the performance of the proposed model on the OD TTE task against the following origin-destination travel time estimation methods.

- **RNE** [49]: determines the path distances between road segments by referencing their latent embeddings.
- **TEMP** [31]: calculates the mean travel times of historical trajectories that align closely in terms of locations and time.
- **LR**: establishes a linear mapping from input features to travel times based on training labels.
- **GBM**: an advanced non-linear regression model, which we implement via XGBoost [50].
- **ST-NN** [32]: concurrently predicts both travel distances and travel times for OD pairs.
- **MURAT** [33]: jointly predicts travel distance and time while utilizing the departure time as extra information.
- **DeepOD** [5]: exploits the correlation between input features and historical trajectories during training.
- **DOT** [22]: a two-stage framework that generates an image representation of trajectories for estimating travel time.

5.2.2 Trajectory Recovery Methods

We compare the proposed model on the TR task against the following trajectory recovery methods, including both non-learning-based and learning-based approaches.

- **Shortest Path** [51]: recovers paths between consecutive points in sparse trajectories with Dijkstra’s shortest path algorithm [52] on road networks.
- **Linear** [53]: extends sparse trajectories using linear interpolation.
- **TrImpute** [23]: imputes sparse trajectories with a crowd wisdom-based algorithm.
- **DHTR** [54]: combines the seq2seq framework [34] and Kalman filtering to recover dense trajectories.
- **AttnMove** [15]: leverages Attention mechanisms to predict the sequence of road segments.
- **MTrajRec** [16]: employs a GRU-based auto-regressive model to recover road network-constrained trajectories.
- **RNTrajRec** [17]: integrates the transformer model and the inherent road network structure to recover trajectories.

5.2.3 Trajectory Prediction Methods

The following trajectory prediction methods for comparison encompass both models specifically designed for the TP task, as well as general trajectory models. To generate predictions, a prediction module implemented with fully-connected network is incorporated into the embedding vectors produced by these general trajectory models.

- **trajectory2vec** [39]: constructs behavior sequences to extract high-level correlations in trajectories.
- **t2vec** [18]: builds upon an auto-encoding framework, offering a more robust general trajectory model.
- **DeepMove** [1]: an end-to-end Attentional RNN-based sequential model designed for predicting future movements.
- **Transformer** [25]: a widely-used sequential model that excels at extracting correlations in long sequences.
- **Trembr** [19]: integrates road network information with auto-encoding for general trajectory modeling.
- **START** [20]: incorporates both spatio-temporal correlations and travel semantics for general trajectory modeling.

It is noteworthy that *Trembr* and *START* require road segment features for optimal functioning, which are not directly obtainable in the context of sparse trajectories. Therefore, the trajectories recovered by the most effective TR baseline, *RNTrajRec*, are served as input to them during evaluation. These combinations are denoted as *Trembr+RNTR* and *START+RNTR*. We also report their performance when they are fed with map-matched, real dense trajectories.

5.2.4 Variations without pre-training or fine-tuning

To further assess the effectiveness of the pre-training and fine-tuning processes within the proposed model, we introduce two additional variations for comparison.

- **UVTM w/o pt**: excludes the pre-training process in Section 4.4, and supervises the model solely with task-specific generation.
- **UVTM w/o ft**: bypasses the fine-tuning process. After pre-training the model, directly employs it to handle tasks with task-specific input arrangements in a zero-shot manner.

5.3 Settings

In both datasets, trajectories are initially sorted by departure time, and then they are divided into training, validation, and testing sets with ratios of 8:1:1. All methods undergo training using the training set, and the validation set is employed for hyper-parameter tuning and early stopping. The final metrics are derived from the testing set. For the TR and TP tasks, three sampling intervals are used, i.e., 1, 2, and 4 minutes, to simulate different levels of sparsity in trajectories. The TR task aims to recover these sparse trajectories into dense ones with a 15-second sampling interval. The TP task aims to predict the road segments, fraction, coordinates, and times of the trajectories’ destinations, given historical trajectories except their last points.

To quantify the performance of the different approaches, we employ different metrics for the three downstream tasks. 1) For the OD TTE task, we use the root mean squared error (RMSE), mean absolute error (MAE), and the mean

absolute percentage error (MAPE) to evaluate the accuracy of estimated travel times. 2) *For the TR task*, following MTrajRec [16], the Precision and Recall of the recovered road segments are determined. Specifically, for each evaluation sample, we gather the set \mathcal{E}_R of recovered road segments and the set \mathcal{E}_G of ground truth road segments from the road segment features. Then, the Precision is calculated as $\frac{|\mathcal{E}_R \cap \mathcal{E}_G|}{|\mathcal{E}_R|}$, and the Recall is calculated as $\frac{|\mathcal{E}_R \cap \mathcal{E}_G|}{|\mathcal{E}_G|}$. Additionally, the MAE of distances for the recovered coordinates and road locations is determined. The distance between a pair of recovered and ground truth trajectory points is determined either by their shortest distance on the Earth's surface or the road network, depending on whether we are evaluating the recovered coordinates or road locations. 3) *For the TP task*, we report the Accuracy of the predicted road segment indices, the distance MAE of the predicted coordinates and road locations, and the MAE of the predicted time. Each experiment is conducted 5 times and the average metric values are reported.

The proposed method is implemented using Python and PyTorch [55]⁴. Baselines are configured according to the optimal parameter settings suggested in their respective papers. Three key hyper-parameters in the proposed model are considered during experiments, with their ranges and optimal values reported in Table 2.

TABLE 2
Hyper-parameter range and optimal values.

Parameter	Range
d	32, 64, <u>128</u> , 192, 256
N_h	1, 2, 4, 8, 16
δ (meters)	10, 50, <u>100</u> , 150, 200

5.4 Comparison with Baselines

5.4.1 Comparison on Overall Accuracy

Tables 3, 4, and 5 present the performance results of different approaches for the OD TTE, TR, and TP tasks, respectively. The results indicate that the proposed model is capable of consistently outperforming comparison methods on diverse trajectory-related tasks, demonstrating its adaptability.

For the OD TTE task, the proposed model performs better than the state-of-the-art baseline, *DOT*, owing to its pre-training that enhances its understanding of the correlations between trajectories and travel times. The superiority is further substantiated by the fact that while *DOT* outperforms *UVTM w/o pt*, it still lags behind the pre-trained and fine-tuned *UVTM*.

For the TR task, a limitation associated with the comparison methods is that they are typically trained to fit sparse trajectories with a specific sampling interval, thereby constraining their generalizability. In contrast, the proposed model, by leveraging pre-training, can adapt to trajectories with varying sampling intervals without necessitating complete re-training. The performance of *UVTM w/o ft* shows that the model can match or even exceed the best-performing baselines without fine-tuning, emphasizing its robustness to varying sampling intervals.

4. Codes available at <https://github.com/Logan-Lin/UVTM>

TABLE 3
Origin-destination travel time estimation accuracy comparison of different approaches.

Datasets	Chengdu / Porto		
Methods	MAE (min) ↓	RMSE (min) ↓	MAPE (%) ↓
RNE	1.087 / 2.357	4.967 / 7.168	18.185 / 53.894
TEMP	0.816 / 2.610	1.100 / 3.414	13.003 / 59.178
LR	0.815 / 2.596	1.097 / 3.408	12.997 / 58.390
GBM	0.773 / 2.200	1.202 / 3.116	11.142 / 43.308
ST-NN	0.770 / 2.136	1.031 / 3.027	12.470 / 45.285
MURAT	0.731 / 1.971	0.979 / 2.827	11.931 / 41.259
DeepOD	0.640 / 1.899	0.880 / 2.780	10.517 / 36.956
DOT	<u>0.614</u> / <u>1.777</u>	<u>0.841</u> / <u>2.644</u>	<u>9.937</u> / <u>34.883</u>
UVTM w/o pt	0.666 / 1.871	0.933 / 2.797	10.501 / 34.895
UVTM w/o ft	2.203 / 3.470	2.469 / 4.694	35.039 / 45.700
UVTM	0.561 / 1.615	0.784 / 2.470	8.853 / 31.391

Bold denotes the best result, underline denotes the second-best result. ↓ means lower is better.

For the TP task, while *Trembr* and *START* are effective in modeling dense trajectories, their performance decreases notably with sparse trajectories. Pairing them with TR methods does not yield optimal results due to the accumulation of errors in this configuration. In contrast, the proposed model is tailored to accommodate sparsity, resulting in excellent performance even when faced with sparse trajectories. Table 6 demonstrates that the proposed model, utilizing sparse trajectories sampled at 1-minute intervals, can even outperform *Trembr* and *START* fed with dense trajectories sampled at 15-second intervals.

5.4.2 Comparison on Efficiency

We assess the efficiency of the different approaches across three dimensions: model size, training time, and testing time. The model size reflects the memory requirements during operation, while the training and testing time provide insights into the efficiency during training and testing. The calculations for model sizes are determined based on the types and numbers of learnable parameters of each approach. Furthermore, the training and testing times are recorded on a machine equipped with an Intel(R) Xeon(R) Gold 5215 CPU and an nVidia(R) Quadro RTX 8000 GPU.

As detailed in Table 7, the proposed model exhibits similar or superior efficiency when compared with the leading methods for each task. Notably, in real-world scenarios where multiple tasks are performed on the same dataset, our model showcases even higher computational efficiency compared to existing solutions. This is because our model can be trained once and perform different tasks, while existing solutions require separate training and storage for each task. For instance, to execute the TTE, TR, and TP tasks on the Chengdu dataset, opting for state-of-the-art baselines would necessitate a training time of 20.675 minutes per epoch. In comparison, training our model only requires 1.668 minutes per epoch.

5.5 Performance Analysis

5.5.1 Efficacy of Pre-training

The advantage of pre-training is assessed by comparing the convergence rate of the proposed model on tasks when

TABLE 4
Trajectory recovery accuracy comparison of different approaches.

Sampling Interval μ		1 minute / 2 minutes / 4 minutes					
Datasets	Methods	Precision (%) \uparrow	Recall (%) \uparrow	MAE (Coor, meters) \downarrow	MAE (Road, meters) \downarrow		
Chengdu	Shortest Path	62.638 / 43.504 / 29.431	59.346 / 40.949 / 27.607	213.10 / 428.69 / 752.19	206.91 / 391.39 / 586.09		
	Linear	66.642 / 48.604 / 36.209	65.557 / 45.234 / 30.496	183.64 / 385.23 / 675.85	169.46 / 378.96 / 564.19		
	TrImpute	77.520 / 60.179 / 57.526	76.202 / 58.461 / 53.747	166.82 / 276.56 / 408.64	155.71 / 265.34 / 387.36		
	DHTR	53.514 / 53.608 / 50.985	58.868 / 47.918 / 46.311	205.59 / 317.45 / 450.61	300.67 / 470.46 / 547.41		
	AttnMove	84.162 / 81.402 / 78.645	81.839 / 76.612 / 69.257	252.59 / 280.20 / 354.39	201.51 / 258.69 / 323.52		
	MTrajRec	85.039 / 82.596 / 80.684	83.351 / 80.113 / 72.952	243.01 / 264.15 / 311.53	173.67 / 204.58 / 282.88		
	RNTrajRec	87.653 / 83.174 / 79.404	86.025 / 80.150 / 72.633	215.24 / 234.27 / 326.92	114.04 / 148.04 / 292.61		
	UVTM w/o pt	83.720 / 77.425 / 72.471	82.827 / 73.933 / 62.757	194.30 / 272.86 / 479.49	86.15 / 230.68 / 448.98		
Porto	UVTM w/o ft	<u>87.664</u> / <u>83.592</u> / <u>79.096</u>	85.837 / 78.406 / 70.717	<u>137.69</u> / <u>217.85</u> / 350.29	<u>81.81</u> / 167.16 / 315.07		
	UVTM	89.071 / 84.373 / 80.828	88.249 / 81.520 / 73.212	133.38 / 192.54 / 304.72	67.98 / 143.48 / 275.37		
	Shortest Path	69.780 / 53.590 / 40.492	60.354 / 46.263 / 33.758	202.17 / 434.37 / 679.72	165.02 / 319.32 / 478.66		
	Linear	72.961 / 60.966 / 48.529	63.146 / 48.401 / 35.507	196.51 / 403.05 / 621.88	132.76 / 275.43 / 430.22		
	TrImpute	76.781 / 66.492 / 50.021	69.599 / 58.676 / 43.052	132.76 / 275.43 / 430.22	128.48 / 235.63 / 347.30		
	DHTR	63.287 / 58.897 / 52.658	62.511 / 56.444 / 42.462	235.32 / 292.65 / 355.23	285.68 / 336.48 / 389.18		
	AttnMove	79.541 / 75.751 / 71.248	67.116 / 56.751 / 48.991	184.70 / 222.51 / 304.31	134.17 / 184.03 / 251.92		
	MTrajRec	78.081 / 72.847 / 64.566	71.853 / 60.068 / 46.110	168.34 / 283.90 / 496.96	121.64 / 215.40 / 391.67		
	RNTrajRec	80.305 / 77.094 / <u>75.573</u>	74.953 / <u>65.370</u> / 50.965	135.17 / <u>175.42</u> / 294.08	111.75 / 152.49 / 230.30		
	UVTM w/o pt	81.058 / 73.842 / 67.189	77.683 / 61.987 / 42.688	132.97 / 229.11 / 496.99	69.70 / 204.01 / 488.13		
	UVTM w/o ft	81.716 / 79.067 / 74.732	75.144 / 63.059 / 50.269	118.79 / 194.37 / 338.30	76.46 / 159.81 / 320.79		
	UVTM	82.640 / 79.494 / 77.068	78.605 / 66.850 / 53.436	103.06 / 164.75 / 282.61	63.18 / 138.66 / 242.08		

Bold denotes the best result, and underline denotes the second-best result. \uparrow means higher is better, and \downarrow means lower is better.

TABLE 5
Trajectory prediction accuracy comparison of different approaches.

Sampling Interval μ		1 minute / 2 minutes / 4 minutes					
Datasets	Methods	Accuracy (%) \uparrow	MAE (Coor, meters) \downarrow	MAE (Road, meters) \downarrow	MAE (Time, seconds) \downarrow		
Chengdu	trajectory2vec	31.496 / 24.403 / 17.163	1514.5 / 1682.8 / 1861.6	1322.0 / 1616.5 / 1957.5	14.474 / 20.722 / 37.256		
	t2vec	53.349 / 43.303 / 35.058	528.65 / 602.45 / 731.00	286.18 / 434.27 / 635.45	13.016 / 19.539 / 34.488		
	DeepMove	58.499 / 45.985 / 37.338	<u>319.14</u> / 461.73 / 664.93	258.96 / 397.95 / 607.42	11.994 / 19.435 / 35.039		
	Transformer	65.192 / 60.028 / <u>55.139</u>	374.36 / <u>402.11</u> / <u>431.01</u>	236.86 / <u>287.61</u> / 320.92	16.287 / 29.848 / 34.226		
	Trembr+RNTR	52.065 / 43.196 / 34.655	421.95 / 482.67 / 561.89	398.76 / 455.98 / 532.03	14.346 / 19.110 / 28.659		
	START+RNTR	59.462 / 48.466 / 40.941	375.39 / 421.45 / 481.32	355.00 / 399.66 / 457.35	12.771 / 14.439 / 19.443		
	UVTM w/o pt	<u>71.795</u> / 50.041 / 33.882	376.05 / 540.89 / 857.43	263.83 / 568.73 / 942.64	<u>5.301</u> / <u>10.198</u> / <u>9.562</u>		
	UVTM w/o ft	67.064 / <u>61.766</u> / 53.949	368.30 / 452.75 / 544.02	<u>230.24</u> / 320.19 / 443.43	6.084 / 12.189 / 9.910		
Porto	UVTM	82.820 / 78.921 / 72.083	260.31 / 303.44 / 362.32	128.44 / 200.80 / 260.13	3.821 / 7.919 / 7.131		
	trajectory2vec	13.396 / 10.178 / 5.440	1709.3 / 2108.9 / 2488.5	2227.4 / 2425.6 / 3003.6	31.442 / 46.097 / 54.585		
	t2vec	38.945 / 30.805 / 22.812	432.77 / 528.89 / 732.13	206.95 / 346.72 / 641.26	17.420 / 28.436 / 48.651		
	DeepMove	43.774 / 33.562 / 23.645	<u>252.22</u> / 390.57 / 679.62	197.89 / 328.93 / 681.40	16.998 / 26.309 / 46.629		
	Transformer	43.441 / 39.425 / 33.685	323.58 / <u>351.18</u> / <u>402.32</u>	216.18 / 256.89 / 283.72	18.231 / 30.541 / 49.083		
	Trembr+RNTR	40.128 / 34.857 / 26.004	413.20 / 471.45 / 620.77	393.18 / 448.74 / 597.73	18.915 / 21.843 / 28.683		
	START+RNTR	52.118 / 43.617 / 34.931	351.98 / 416.83 / 503.32	333.56 / 396.39 / 483.41	14.729 / 17.722 / 22.455		
	UVTM w/o pt	<u>61.480</u> / 46.938 / 31.304	307.23 / 369.86 / 649.71	<u>145.36</u> / 295.58 / 624.11	<u>10.853</u> / <u>11.316</u> / <u>11.643</u>		
	UVTM w/o ft	48.770 / <u>48.474</u> / <u>44.477</u>	325.58 / 370.04 / 454.70	170.32 / <u>193.40</u> / <u>279.12</u>	11.509 / 18.342 / 18.081		
	UVTM	68.529 / 67.697 / 65.125	238.83 / 232.15 / 322.63	99.67 / 103.79 / 140.17	6.845 / 9.449 / 11.549		

Bold denotes the best result, and underline denotes the second-best result. \uparrow means higher is better, and \downarrow means lower is better.

TABLE 6
Trajectory prediction accuracy comparison between baselines fed with dense trajectories and the proposed model.

Datasets	μ	Accuracy (%)	MAE (Coor, m.)	MAE (Time, sec.)
Trembr	15 sec.	66.077 / 52.063	399.40 / 362.84	8.534 / 15.046
START	15 sec.	<u>74.990</u> / 62.997	<u>354.76</u> / 318.26	8.068 / 12.476
UVTM	4 min.	72.083 / 65.125	362.32 / 322.63	<u>7.131</u> / 11.549
UVTM	1 min.	82.820 / 68.529	260.31 / 238.83	3.821 / 6.845

employed with and without pre-training. Specifically, we track the performance metrics of the proposed model on tasks across training epochs. The findings are shown in Figure 7.

The observations are clear: across both TR and TP tasks, the pre-trained model consistently outperforms its non-pre-trained counterpart in terms of convergence rate. This demonstrates the model's capability to smoothly transition between different tasks with minimal fine-tuning. By achieving superior performance in fewer epochs, the pre-training process enhances computational efficiency, partic-

Bold denotes the best result, underline denotes the second-best result.

TABLE 7
Efficiency metrics of different approaches.

Datasets		Chengdu / Porto		
Tasks	Methods	Model size (MBytes)	Train time (min/epoch)	Test time (sec)
OD TTE	RNE	2.446 / 2.173	0.100 / 0.040	0.170 / 0.062
	ST-NN	1.185 / 1.185	0.112 / 0.082	0.220 / 0.085
	MURAT	9.120 / 8.847	0.153 / 0.095	0.210 / 0.075
	DeepOD	8.184 / 7.928	0.382 / 0.171	0.328 / 0.099
	DOT	8.763 / 8.496	1.552 / 0.752	1.672 / 0.926
	UVTM	10.146 / 9.333	0.533 / 0.336	1.231 / 0.647
TR	TrImpute	2.778 / 1.262	- / -	6.110K / 2.199K
	DHTR	6.426 / 6.426	0.176 / 0.115	2.503K / 0.257K
	AttnMove	6.799 / 6.250	6.844 / 2.460	92.673 / 32.081
	MTrajRec	19.180 / 18.495	8.470 / 4.437	0.125K / 54.062
	RNTrajRec	20.639 / 19.876	8.925 / 4.463	0.144K / 60.861
	UVTM	10.146 / 9.333	1.668 / 1.470	21.984 / 15.237
TP	trajectory2vec	6.306 / 6.306	0.158 / 0.085	0.204 / 0.110
	t2vec	7.170 / 7.170	0.278 / 0.109	0.221 / 0.113
	DeepMove	5.282 / 5.282	0.253 / 0.126	0.567 / 0.292
	Transformer	6.295 / 6.295	1.090 / 0.538	3.568 / 1.596
	Trembr+RNTR	26.103 / 25.792	9.468 / 5.072	0.149K / 62.500
	START+RNTR	28.708 / 27.099	10.198 / 5.420	0.158K / 65.468
	UVTM	10.146 / 9.333	1.667 / 1.450	3.357 / 1.703

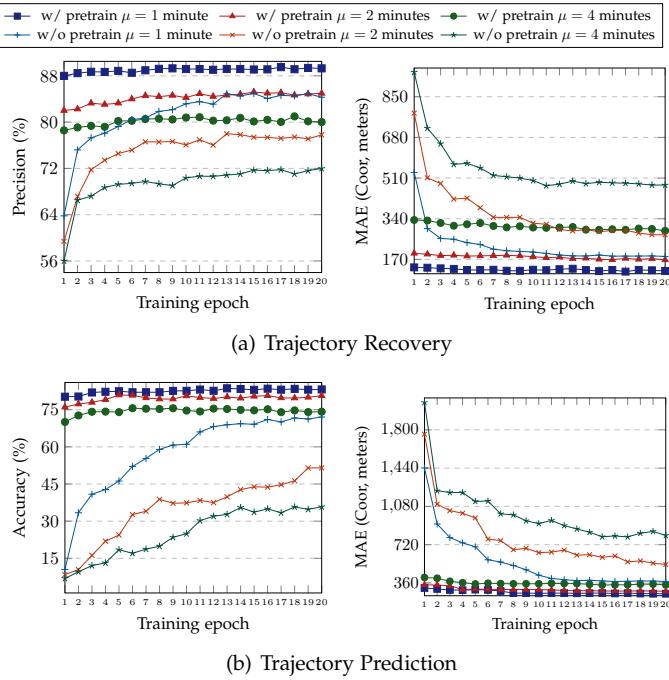


Fig. 7. Comparison of convergence rate with or without pre-training in Chengdu.

ularly in situations where a single dataset is utilized for multiple tasks.

5.5.2 Scalability of Pre-training with Limited Data

The construction of pre-training samples in the proposed model requires a certain amount of densely sampled trajectories. Thus, it is of interest to determine how effectively the model operates when faced with limited dense trajectories. To assess this aspect, we examine the scalability of the model in scenarios where only a subset of the training dense trajectories is available for the pre-training phase.

The results in Figure 8 offer several insights. Notably, the proposed model exhibits strong consistency. Even when the pre-training is performed with limited data, the per-

formance remains relatively robust. This underscores the model's potential in practical contexts where obtaining large-scale dense trajectories might pose challenges. Furthermore, a significant improvement in performance is evident when comparing the results at 20% scale to those of the model without any pre-training (0% scale). This is further evidence of the benefits of the pre-training.

5.5.3 Scalability of Fine-tuning with Limited Task-specific Data

Comparing the model's performance with and without fine-tuning in both Tables 3 and 5, reveals a clear advantage when fine-tuning is applied. This can be credited to the difference in input arrangements between pre-training and specific tasks. Such a marked difference raises the question: does the model rely on extensive fine-tuning datasets to achieve optimal performance?

To answer the question, we investigate the model's scalability under scenarios where only a fraction of the task-specific data is available for fine-tuning. It is clear from the results in Figure 9 that the proposed model exhibits a remarkable stability at approximating its best performance when given just 20% of the full fine-tuning datasets. Comparing these results against a scenario where no fine-tuning is performed (0% scale), we see that even a minuscule fine-tuning dataset can facilitate the model to improve its accuracy on tasks. These findings underscore the model's utility in the real-world, where gathering large-scale task-specific labeled datasets can be challenging.

5.5.4 Effectiveness of Hyper-parameters

We analyze the influence of the three primary hyper-parameters in Table 2 on the performance of the model. The results are presented in Figure 10. We highlight the following observations:

- 1) The embedding dimensionality d determines the expressive power of the model. As evident from Figure 10(a), an increase in d typically yields improved performance. However, beyond $d = 128$, the gains in performance are marginal, while the computational and memory overheads continue to grow. Therefore, $d = 128$ appears to be an ideal trade-off between performance and computational efficiency.
- 2) The number of Attention heads N_h controls the representational capacity of the encoder in our model. Drawing from Figure 10(b), the most beneficial value of both Precision and MAE occur at $N_h = 8$.
- 3) The distance threshold δ determines the spatial scope of road segment neighbors. The insights from Figure 10(c) suggest its primary effect is on the precision of the generated road segments, with minimal influence on the accuracy of predicted coordinates. A peak in Precision is observed at $\delta = 100$. Employing a smaller threshold might result in missed road segment candidates, while a larger threshold can introduce more noise into the generation progress.

5.5.5 Effectiveness of Modules

To determine the impact of individual features and modules in the proposed model, we conduct an ablation study. This

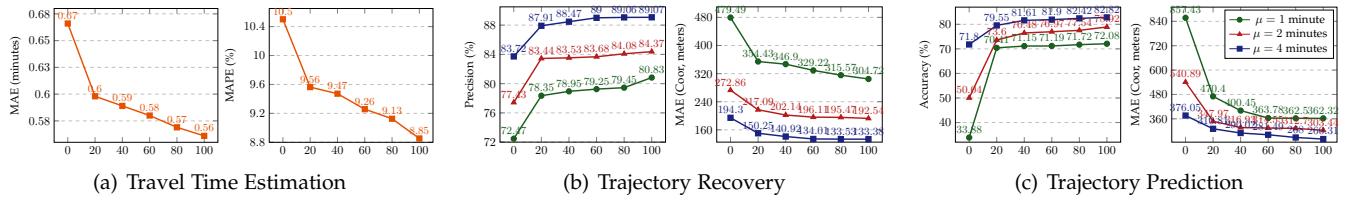


Fig. 8. Scalability with regard to the size (%) of pre-train sets in Chengdu.

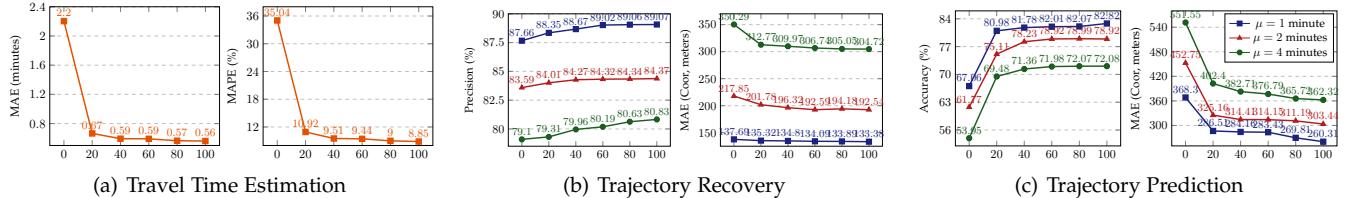


Fig. 9. Scalability with regard to the size (%) of fine-tune sets in Chengdu.

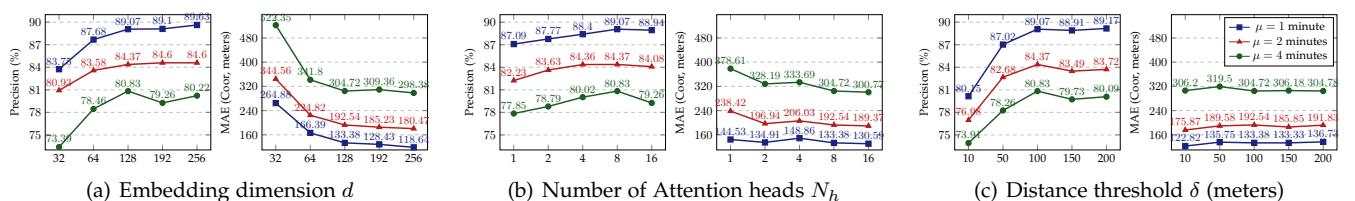


Fig. 10. Effectiveness of hyper-parameters on the trajectory recovery task in Chengdu.

TABLE 8

Effectiveness of features and modules on the sparse trajectory recovery task in Chengdu.

Sampling Interval μ	1 minute / 2 minutes / 4 minutes		
	Variations	Precision (%)	MAE (Coor, meters)
w/o neigh.	76.834 / 74.494 / 71.950	128.36 / 193.41 / 300.30	
w/o coor.	78.891 / 77.345 / 72.433	162.03 / 284.03 / 444.69	
w/o time	86.821 / 83.230 / 77.055	140.98 / 203.23 / 318.81	
w/o shuffle	84.406 / 77.467 / 73.589	158.95 / 276.26 / 433.86	
Flat encoder	85.548 / 79.874 / 74.162	142.02 / 216.87 / 323.66	
FC num. enc.	86.051 / 81.433 / 76.964	145.67 / 218.25 / 327.71	
UVTM	89.071 / 84.343 / 80.828	133.38 / 192.54 / 304.72	

involves contrasting the performance of the complete model with the performance of the following variations.

- 1) *w/o neigh.*: excludes the set of road segment neighbors $\Omega(l_i)$ in Equation 1.
- 2) *w/o coor.*: excludes the coordinate feature l_i in Equation 1.
- 3) *w/o time*: excludes the time feature t_i in Equation 1.
- 4) *w/o shuffle*: does not shuffle the order of generated spans during the pre-training procedure.
- 5) *Flat encoder*: calculates h_i as the mean pooling of Z_i instead of using self-attention.
- 6) *FC num. enc.*: replaces the encoding module Φ with a fully connected network.

The results are computed on Chengdu's test set for the TR task, as shown in Table 8. We draw the following observations:

- 1) The road segment neighbors, coordinates, and time features collectively enhance the efficacy of the model. Ex-

cluding any of these features yields a noticeable drop in performance.

- 2) Omitting shuffling when performing pre-training limits the model's ability to capture bi-directional correlations in trajectories, causing a decline in performance.
- 3) The hierarchical attention in the trajectory encoder and the encoding module in the feature domain embedding layer are effective at enhancing the model's performance.

6 CONCLUSION

In this paper, we introduce UVTM, a versatile and robust universal vehicle trajectory model. UVTM is distinguished by its ability to be trained once on a given dataset and subsequently adapt to a variety of trajectory-related tasks without relying on additional prediction modules. To ensure its applicability across diverse tasks, particularly those involving input of incomplete trajectories, UVTM separates spatio-temporal features in trajectories into three distinct domains. This design enables the independent masking and generation of each domain to accommodate the unique input and output needs of different tasks. Furthermore, UVTM's robustness to sparsity is enhanced through a pre-training process. During this process, UVTM learns to reconstruct the feature domains of densely sampled trajectories from their re-sampled sparse counterparts, thus maintaining its performance in the presence of sparsity. The versatility and robustness of UVTM are able to improve both computational and storage efficiency in ITS, facilitating the execution of multiple tasks with a single model and enabling efficient dataset downsizing through re-sampling. Empirical

evaluation on two real-world vehicle trajectory datasets and three representative tasks demonstrates UVTM's superior performance, proving UVTM's effectiveness in achieving its objective of universal trajectory modeling.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No. 62272033).

REFERENCES

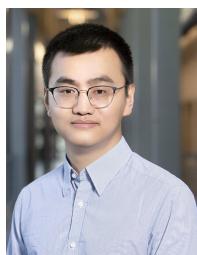
- [1] J. Feng, Y. Li, C. Zhang, F. Sun, F. Meng, A. Guo, and D. Jin, "Deepmove: Predicting human mobility with attentional recurrent networks," in *WWW*, 2018, pp. 1459–1468.
- [2] D. Kong and F. Wu, "HST-LSTM: A hierarchical spatial-temporal long-short term memory network for location prediction," in *IJCAI*, 2018, pp. 2341–2347.
- [3] Y. Liang and Z. Zhao, "Netraj: A network-based vehicle trajectory prediction model with directional representation and spatiotemporal attention mechanisms," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 14470–14481, 2022.
- [4] Z. Fang, L. Pan, L. Chen, Y. Du, and Y. Gao, "MDTP: A multi-source deep traffic prediction framework over spatio-temporal trajectory data," *Proc. VLDB Endow.*, vol. 14, no. 8, pp. 1289–1297, 2021.
- [5] H. Yuan, G. Li, Z. Bao, and L. Feng, "Effective travel time estimation: When historical trajectories over road networks matter," in *SIGMOD*, 2020, pp. 2135–2149.
- [6] K. Fu, F. Meng, J. Ye, and Z. Wang, "Compacteta: A fast inference system for travel time prediction," in *KDD*, 2020, pp. 3337–3345.
- [7] C. Guo, B. Yang, J. Hu, C. S. Jensen, and L. Chen, "Context-aware, preference-based vehicle routing," *VLDB J.*, vol. 29, no. 5, pp. 1149–1170, 2020.
- [8] S. A. Pedersen, B. Yang, and C. S. Jensen, "Anytime stochastic routing with hybrid learning," *Proc. VLDB Endow.*, vol. 13, no. 9, pp. 1555–1567, 2020.
- [9] Y. Liu, K. Zhao, G. Cong, and Z. Bao, "Online anomalous trajectory detection with deep generative sequence modeling," in *ICDE*, 2020, pp. 949–960.
- [10] X. Han, R. Cheng, C. Ma, and T. Grubenmann, "Deepte: Effective and efficient online time-dependent trajectory outlier detection," *Proc. VLDB Endow.*, vol. 15, no. 7, pp. 1493–1505, 2022.
- [11] M. Yue, Y. Li, H. Yang, R. Ahuja, Y. Chiang, and C. Shahabi, "DETECT: deep trajectory clustering for mobility-behavior analysis," in *IEEE BigData*, 2019, pp. 988–997.
- [12] T. Li, L. Chen, C. S. Jensen, T. B. Pedersen, Y. Gao, and J. Hu, "Evolutionary clustering of moving objects," in *ICDE*, 2022, pp. 2399–2411.
- [13] S. B. Yang, C. Guo, and B. Yang, "Context-aware path ranking in road networks," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 7, pp. 3153–3168, 2022.
- [14] Z. Fang, S. Gong, L. Chen, J. Xu, Y. Gao, and C. S. Jensen, "Ghost: A general framework for high-performance online similarity queries over distributed trajectory streams," *Proc. ACM Manag. Data*, vol. 1, no. 2, pp. 173:1–173:25, 2023.
- [15] T. Xia, Y. Qi, J. Feng, F. Xu, F. Sun, D. Guo, and Y. Li, "Attnmove: History enhanced trajectory recovery via attentional network," in *AAAI*, 2021, pp. 4494–4502.
- [16] H. Ren, S. Ruan, Y. Li, J. Bao, C. Meng, R. Li, and Y. Zheng, "Mtrajrec: Map-constrained trajectory recovery via seq2seq multi-task learning," in *KDD*, 2021, pp. 1410–1419.
- [17] Y. Chen, H. Zhang, W. Sun, and B. Zheng, "Rntrajrec: Road network enhanced trajectory recovery with spatial-temporal transformer," pp. 829–842, 2023.
- [18] X. Li, K. Zhao, G. Cong, C. S. Jensen, and W. Wei, "Deep representation learning for trajectory similarity computation," in *ICDE*, 2018, pp. 617–628.
- [19] T. Fu and W. Lee, "Trembr: Exploring road networks for trajectory representation learning," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 1, pp. 10:1–10:25, 2020.
- [20] J. Jiang, D. Pan, H. Ren, X. Jiang, C. Li, and J. Wang, "Self-supervised trajectory representation learning with temporal regularities and travel semantics," *CoRR*, vol. abs/2211.09510, 2022.
- [21] S. B. Yang, J. Hu, C. Guo, B. Yang, and C. S. Jensen, "Lightpath: Lightweight and scalable path representation learning," in *KDD*, 2023, pp. 2999–3010.
- [22] Y. Lin, H. Wan, J. Hu, S. Guo, B. Yang, Y. Lin, and C. S. Jensen, "Origin-destination travel time oracle for map-based services," *SIGMOD*, vol. 1, no. 3, pp. 1–27, 2023.
- [23] M. M. Elshrif, K. Isufaj, and M. F. Mokbel, "Network-less trajectory imputation," in *SIGSPATIAL*, 2022, pp. 8:1–8:10.
- [24] F. Zhou, Y. Dai, Q. Gao, P. Wang, and T. Zhong, "Self-supervised human mobility learning for next location prediction and trajectory classification," *Knowl. Based Syst.*, vol. 228, p. 107214, 2021.
- [25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.
- [26] Z. Wang, K. Fu, and J. Ye, "Learning to estimate the travel time," in *KDD*, 2018, pp. 858–866.
- [27] D. Wang, J. Zhang, W. Cao, J. Li, and Y. Zheng, "When will you arrive? estimating travel time based on deep neural networks," in *AAAI*, vol. 32, no. 1, 2018, pp. 2500–2507.
- [28] F. Wu and L. Wu, "Deepeta: a spatial-temporal sequential neural network model for estimating time of arrival in package delivery system," in *AAAI*, vol. 33, no. 01, 2019, pp. 774–781.
- [29] Y. Gan, H. Zhang, and M. Wang, "Travel time estimation based on neural network with auxiliary loss," in *SIGSPATIAL*, 2021, pp. 642–645.
- [30] L. Yang, S. Jiang, and F. Zhang, "Multitask learning with graph neural network for travel time estimation," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.
- [31] H. Wang, X. Tang, Y.-H. Kuo, D. Kifer, and Z. Li, "A simple baseline for travel time estimation using large-scale trip data," *ACM Trans. on Intelli. Sys. and Tech.*, vol. 10, no. 2, pp. 1–22, 2019.
- [32] I. Jindal, Z. T. Qin, X. Chen, M. S. Nokleby, and J. Ye, "A unified neural network approach for estimating travel time and distance for a taxi trip," *CoRR*, vol. abs/1710.04350, 2017.
- [33] Y. Li, K. Fu, Z. Wang, C. Shahabi, J. Ye, and Y. Liu, "Multi-task representation learning for travel time estimation," in *KDD*, 2018, pp. 1695–1704.
- [34] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *NeurIPS*, 2014, pp. 3104–3112.
- [35] C. Miao, Z. Luo, F. Zeng, and J. Wang, "Predicting human mobility via attentive convolutional network," in *WSDM*, 2020, pp. 438–446.
- [36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [37] B. Yan, G. Zhao, L. Song, Y. Yu, and J. Dong, "Precln: Pretrained-based contrastive learning network for vehicle trajectory prediction," *WWW*, pp. 1–23, 2022.
- [38] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, vol. 119, 2020, pp. 1597–1607.
- [39] D. Yao, C. Zhang, Z. Zhu, J. Huang, and J. Bi, "Trajectory clustering via deep representation learning," in *IJCNN*, 2017, pp. 3880–3887.
- [40] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [41] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018.
- [42] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019, pp. 4171–4186.
- [43] P. Chao, Y. Xu, W. Hua, and X. Zhou, "A survey on map-matching algorithms," in *ADC*, ser. Lecture Notes in Computer Science, vol. 12008, 2020, pp. 121–133.
- [44] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang, "GLM: general language model pretraining with autoregressive blank infilling," in *ACL*, 2022, pp. 320–335.
- [45] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," *NeurIPS*, vol. 33, pp. 7537–7547, 2020.
- [46] Y. Li, S. Si, G. Li, C.-J. Hsieh, and S. Bengio, "Learnable fourier features for multi-dimensional spatial positional encoding," *NeurIPS*, vol. 34, pp. 15816–15829, 2021.
- [47] C. Yang and G. Gidofalvi, "Fast map matching, an algorithm integrating hidden markov model with precomputation," *Int. Journal of Geographical Information Science*, vol. 32, no. 3, pp. 547–570, 2018.

- [48] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *NeurIPS*, 2019, pp. 5754–5764.
- [49] S. Huang, Y. Wang, T. Zhao, and G. Li, "A learning-based method for computing shortest path distances on road networks," in *ICDE*, 2021, pp. 360–371.
- [50] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *KDD*, 2016, pp. 785–794.
- [51] E. W. Chambers, B. T. Fasy, Y. Wang, and C. Wenk, "Map-matching using shortest paths," *ACM Trans. Spatial Algorithms Syst.*, vol. 6, no. 1, pp. 6:1–6:17, 2020.
- [52] D. B. Johnson, "A note on dijkstra's shortest path algorithm," *Journal of the ACM*, vol. 20, no. 3, pp. 385–388, 1973.
- [53] S. Hoteit, S. Secci, S. Sobolevsky, C. Ratti, and G. Pujolle, "Estimating human trajectories and hotspots through mobile phone data," *Comput. Networks*, vol. 64, pp. 296–307, 2014.
- [54] J. Wang, N. Wu, X. Lu, W. X. Zhao, and K. Feng, "Deep trajectory recovery with fine-grained calibration using kalman filter," *IEEE Trans. on Know. and Data Eng.*, vol. 33, no. 3, pp. 921–934, 2019.
- [55] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *NeurIPS*, 2019, pp. 8024–8035.



Yan Lin received the B.S. degree in computer science from Beijing Jiaotong University, Beijing, China, in 2019.

He is currently working toward the Ph.D. degree in the School of Computer and Information Technology, Beijing Jiaotong University. His research interests include spatio-temporal data mining and representation learning.



Jilin Hu received the Ph.D. degree in computer science from Aalborg University, Aalborg, Denmark in 2019.

He is a Professor at the School of Data Science and Engineering, East China Normal University. His research interests include spatio-temporal data analytics and transportation data mining.



Shengnan Guo received the Ph.D. degree in computer science from Beijing Jiaotong University, Beijing, China, in 2021.

She is an associate professor at the School of Computer and Information Technology, Beijing Jiaotong University. Her research interests focus on spatial-temporal data mining and intelligent transportation systems.



Bin Yang received the Ph.D. degree in computer science from Fudan University, Shanghai, China in 2010.

He is a Professor at the School of Data Science and Engineering, East China Normal University. His research interests include spatio-temporal data analytics, machine learning, and data management.



Christian S. Jensen received the Ph.D. degree from Aalborg University in 1991 after 2 1/2 years of study at University of Maryland, and he received the Dr.Techn. degree from Aalborg University in 2000.

He is a Professor at the Department of Computer Science, Aalborg University. His research concerns primarily temporal and spatio-temporal data management and analytics, including indexing and query processing, data mining, and machine learning.



Youfang Lin received the Ph.D. degree in signal and information processing from Beijing Jiaotong University, Beijing, China, in 2003.

He is a Professor with the School of Computer and Information Technology, Beijing Jiaotong University. His main fields of expertise and current research interests include big data technology, intelligent systems, complex networks, and traffic data mining.



Huaiyu Wan received the Ph.D. degree in computer science and technology from Beijing Jiaotong University, Beijing, China, in 2012.

He is a Professor with the School of Computer and Information Technology, Beijing Jiaotong University. His current research interests focus on spatio-temporal data mining, social network mining, information extraction, and knowledge graph.